DOCUMENT RESUME

ED 426 077 TM 029 301

AUTHOR Hickey, Daniel T.; Wolfe, Edward W.; Kindfield, Ann C. H.
TITLE Assessing Learning in a Technology-Supported Genetics Environment: Evidential and Systemic Validity Issues.
SPONS AGENCY National Science Foundation, Arlington, VA.
PUB DATE 1998-04-00
NOTE 51p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
CONTRACT RED-95-53438
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Computer Assisted Testing; Curriculum Development; *Educational Technology; Evaluation Methods; *Genetics; *High School Students; High Schools; Item Response Theory; Qualitative Research; *Student Evaluation; Teaching Methods; *Validity
IDENTIFIERS Consequential Evaluation; Rasch Model

ABSTRACT

To evaluate student learning in a computer-supported environment known as "GenScope," a system was developed for assessing students' understanding and learning of introductory genetics material presented in two developed GenScope instruments. Both quantitative and qualitative methods were used to address traditional evidential validity concerns as well as more contemporary concerns with consequential and systemic validity. Findings from three GenScope implementation classrooms and interviews with two teachers and five secondary school students show strong evidential validity, but only limited consequential validity. In response to these findings, a set of curricular activities was developed to scaffold student assessment performance without compromising the evidential validity of the assessment system. The study shows the usefulness of newer interpretive models of validity inquiry and the value of multifaceted Rasch measurement tools for conducting such inquiry. Two appendixes contain sample items from one assessment and a sample GenScope investigation. (Contains 3 tables, 5 figures, and 23 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 426 077

Assessing Learning in a Technology-Supported Genetics Environment: Evidential and Systemic Validity Issues

Daniel T. Hickey, Georgia State University
Edward W. Wolfe, University of Florida
Ann C. H. Kindfield, Montclair State University

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Daniel T. Hickey
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

April, 1998

Presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Author Notes

This research was partly supported by a Postdoctoral Fellowship from Educational Testing Service and was initiated when all three authors were affiliated with the Center for Performance Assessment at ETS. The *GenScope* Assessment Project is funded by the National Science Foundation (Project # RED-95-53438), via The Concord Consortium. We gratefully acknowledge the input of colleagues, including Joan Heller, Carol Myford, and Drew Gitomer, Bob Mislevy, and Iris Tabak at ETS, Paul Horwitz, Mary Ann Christie, and Joyce Schwartz at the Concord Consortium, and Alex Heidenberg at Georgia State University. Thanks also to the teachers and students at Lincoln-Sudbury High and Boston High. For information, contact: Daniel T. Hickey, Dept EPSE, Georgia State University, Atlanta, GA, 30303, or dhickey@gsu.edu

TM029301

Abstract

In order to evaluate student learning in a computer-supported environment known as *GenScope*, we developed a system for assessing students' understanding and learning of introductory genetics. A critical aspect of the development effort concerned the *validity* of this assessment system. We used quantitative and qualitative methods to address traditional evidential validity concerns as well as more contemporary concerns with *consequential* and *systemic* validity. Specifically, we examined whether or not our assessment system helped students develop the understanding it was designed to assess. Our inquiry revealed strong evidential validity, but only limited consequential validity. In response we developed a set of curricular activities designed to scaffold student assessment performance without compromising the evidential validity of the assessment system. In addition to documenting and enhancing the system's validity, these efforts demonstrate the utility of newer interpretive models of validity inquiry and the value of mutifaceted Rasch measurement tools for conducting such inquiry.

Assessing Learning in a Technology-Supported Genetics Environment
Evidential and Systemic Validity Issues

The work described here was a result of our participation in a multi-year implementation and evaluation effort involving a computer-supported learning environment known as *GenScope* (Horwitz, Neumann, & Schwarz, 1996, see http //GenScope concord org) GenScope was designed primarily for teaching introductory genetics in secondary Biology classrooms As illustrated in Figure 1, the GenScope software employs fanciful species such as dragons as well as real species, and lets students observe and manipulate the dynamic relationships across the various levels of biological organization In key respects, this application of educational computing is consistent with recent policy recommendations for K-12 educational technology issued by the President's Committee of Advisors on Science and Technology (PCAST, 1997) Specifically, the software and associated curriculum were designed to help students develop the kind of higher-level domain reasoning skills called for by current science education standards (e g , National Research Council, 1996) and embody contemporary constructivist pedagogical principles

A key challenge in our research effort was developing an assessment system for documenting the degree to which students could demonstrate the kinds of domain reasoning that GenScope ostensibly affords We needed an assessment system that was consistent with the pedagogical assumptions embodied in GenScope while also affording the sort of rigorous evaluation of learning outcomes that are also called for in current policy recommendations (i e , PCAST, 1997) A major part of this challenge—and the focus of this paper—concerns the *validity* of this assessment system This paper describes the assessment system that we developed and our inquiry into its validity This inquiry used interpretive and empirical methods to address traditional evidential validity concerns as well as more contemporary concerns with *consequential* validity (Messick, 1989) and *systemic* validity (Frederiksen & Collins, 1989) In particular, our inquiry considered whether the assessment system further contributed to student learning, and whether or not it had done so at the expense of the system's evidential validity

Validity Inquiry as Argumentation

Assuming that psychological conclusions (including ones about validity) are at some level undetermined (i.e., subject to multiple interpretations), one must marshal evidence in favor of one's own reasoned interpretation and against alternative interpretations. A prerequisite to scientific argumentation is setting forth the assumptions and guiding conceptualizations of the world in which the argument takes place. Overarching assumptions, in the form of "world views" provide communities of scientists with the shared lore of what "counts" as evidence--specifically what constitutes legitimate research questions, acceptable experiments to test those questions, and legitimate data from those experiments. Following are the assumptions about knowing and learning, transfer, assessment, and validity that guided our research and frame the arguments that warrant our conclusions.

Assumptions about Knowing and Learning

Socio-constructivist/situative epistemological perspectives such as *situated cognition* (e.g., Brown, Collins, & Duguid, 1989) hold that knowledge and skills are fundamentally contextualized (i.e., "situated") in the physical and social context in which they are acquired and used. Skills and knowledge are conceptualized as being distributed across the social and physical environment, jointly composed in a system that comprises an individual and peers, teachers, and culturally provided tools. From this perspective, complex cognitive performances usually require external tools, such as pencil, paper, computers, books, peers, teachers, etc. Furthermore, people with less education and skill rely more on these tools for complex thinking than their more proficient counterparts. Rather than something that can be "possessed", proficiency in a domain is seen in part as knowing how to overcome the limits of mind and memory.

Technology-based tools such as GenScope make complex relationships and interactions in particular domains visible and manipulable, allowing students to test their ideas and understanding. Akin to the Cuisinare rods that have dramatically reshaped primary mathematics instruction, the various windows in GenScope provide manipulable representations of a simplified genetic system of imaginary and real organisms. Because the representations are dynamically linked, students can control the abstract processes and observe the components across the various "levels" of biological organization where genetics is manifested (i.e., DNA, chromosomal, cellular, Mendelian, and evolutionary). From traditional empiricist/associationist

BEST COPY AVAILABLE

or pragmatist/rationalist epistemological perspectives, GenScope would be seen as new tool to teach genetics via demonstrations and routine exercises, or as a discovery learning environment in which students can "discover" important domain concepts. In contrast, contemporary constructivist perspectives view GenScope as a tool that affords a structured environment where learners collaboratively experience, learn, and demonstrate a more sophisticated understanding of complex relationships and phenomena than they could otherwise. From this perspective, students initially "understand" the domain represented by a simulation as they internalize the language, representations, and relations in that environment. This internalization happens both as students interact with the environment, and as students interact with each other within that environment. Consider, for example, two students using GenScope to solve complex inheritance problems (e.g., sex linkage or crossover) while struggling with fragile understandings of the underlying concepts such as chromosome type and meiotic events. The common representation afforded by their shared understanding of the simulation environments supports a more sophisticated level of interactions than would be possible without such a tool. Meanwhile, the associated curricular activities and the teacher help these students connect their shared activity to the broader domain (textbook depictions of genetics, other biological domains, other sciences, etc.). As individuals internalize the shared understanding of concepts and phenomenon that are "stretched across" this environment, they move closer and closer to the goal of "expert" understanding in the domain. This exemplifies precisely how contemporary instructional theorists believe that software tools can facilitate learning by extending and expanding what Vygotsky (1978) characterized as "the zone of proximal development"

Assumptions about Transfer

A key aspect of any interpretation of assessment performance is whether it demonstrates *transfer* of knowledge from the learning situation represented by particular learning environment. In the typical absence of a known transfer situation (such as an employment setting or a subsequent course), the actual transfer situation is unknown, and the assessments themselves are the transfer situation. Thus, the validity of one's interpretation of student performance is partly contingent on the appropriateness of the assessments as criteria of performance itself, or as surrogates for some other unspecified transfer setting

From a situated cognition perspective, transfer is considered in terms of the constraints and affordances that support activity in the learning situation and in the transfer situation (Greeno, Collins, & Resnick, 1996). Analyzing transfer involves analyzing the "transformations" that relate a given pair of learning and transfer situations. For any transfer, some constraints and affordances must be the same (be "invariant") across both situations. If transfer is to take place, the learner must learn (become "attuned" to) these *invariants* in the initial learning situation. In order to interpret the degree of transfer represented by performance on assessments, one must first identify the dimensions that vary between the learning situation (i.e., the GenScope environment or a comparison genetics learning environment) and the transfer situation (i.e., our various assessment tasks). For example, one dimension of transfer in our research concerned the way the organism's genotype was represented. As illustrated in Figure 1, GenScope's chromosome window provides a colorful depiction of the organisms' various allelic combinations (i.e., *AA* vs. *Aa* vs. *aa*) that is dynamically linked to other representations, in contrast, our assessments used the traditional "stick figure" representation of the organism's genome. If students' understanding of genotypic representation is to transfer from the GenScope environment to the assessment environment, they must be able to distinguish between the aspects of the representation that are particular to GenScope and the aspects that are invariant (i.e., the domain-relevant information that is conveyed by both representations).

Assumptions about Assessment.

Traditional assessment approaches that tested whether or not an individual "possessed" proficiency were premised on two key assumptions--that knowledge can be decomposed into elements, and that knowledge can be decontextualized in a manner that it can exist or be measured free of context. The perspectives on knowing and learning described above have led many contemporary theorists to reject both assumptions, offering critical implications for assessment.

> Any individual has a range of knowledge and competencies, rather than some fixed level of performance. Depending on how much support and familiarity with the materials at hand she or he has, an individual's performance will be greatly affected. It may be just as

> crucial to measure the quality of that supported performance--or the gap between solo and supported thought (Wolfe, Bixby, Glenn, and Gardner, 1991, p 51)

Many of the conflicts that emerge when developing assessment frameworks are rooted in conflicting assumptions about knowing and learning, and the implications of those assumptions for our assumptions about proficiency and transfer For example, a critical issue concerns the difficulty of the assessments relative to the students' abilities Learning environments such as GenScope are designed to focus on "higher-order" domain-specific understanding, rather than mere memorization of terms and understanding of simple concepts Higher-order thinking is generally characterized as non-algorithmic, complex, and effortful, and as involving multiple solutions, nuanced judgment, the application of multiple criteria, and self-regulation In our development effort, we found that the problems we first designed to assess what we defined as higher-order domain understanding were exceedingly difficult for participants in the initial pilot implementation of the learning environment There are arguments both for and against targeting a level of performance that few students are likely to attain Current perspectives on assessment suggest that a more fundamental question concerns whether or not we have identified an appropriate performance criterion, or whether or not it is appropriate to even select a criterion As illustrated by Wolfe et al above, a central theme of new assessment perspectives is that assessments should maximize student performance as much as possible, starting at whatever level of performance students are capable of This is particularly the case if an intended consequence of assessment is increasing student understanding For example, Wiggins argues

> To make tests truly enabling we must do more than just couch test tasks in more authentic performance contexts We must construct tests that assess whether students are learning how to learn, *given what they know* Instead of testing whether students have learned how to read, we should test their ability to read to learn, instead of finding out whether they "know" formulas, we should find out whether they can use formulas to find other formulas (1993, p 214, emphasis added)

These perspective can be seen as arguing against specifying an "adequate" criterion level of performance At a minimum, this perspective suggests that if a criterion is used, it should

concern the degree of support needed for students to perform at an acceptable level, rather than a level of performance to be achieved without support. Interpretations of student performance can then be made in terms of the type and degree of support needed to solve problems that require "higher-order understanding". Thus, the range of proficiency might start with a highly scaffolded problem, with increasingly higher levels of proficiency indicated by solving the similar problems after stripping away more and more layers of support. As we will show, this perspective (and our initial experience with unscaffolded higher-level problems) led us to design assessments where the easier activities that students first encounter scaffold their performance on the more difficult ones that appear later in the test.

Assumptions about Validity

In one oft-cited characterization, Messick defines validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (1989, p. 5). Messick (e.g., 1989, 1995) discusses validity in terms of the four "facets" derived from crossing the distinction between *interpreting* and *using* test scores, and the distinction between the *evidential* basis of validity and the *consequential* basis of validity (shown in Table 1). Facet 1 is best understood as the search for *construct irrelevant variance*. For example, do students perform better or worse on an assessments for reasons other than individual differences in the underlying targeted construct? Within this explicitly additive framework, Facet 2 adds to Facet 1 the need for evidence that supports the relevance of a given score interpretation in a particular applied setting. For example, is one's presumably valid interpretation of student understanding itself a valid means of assessing the particular learning environment?

The inclusion of the consequences of test use and test score interpretation represents a recent, somewhat controversial advance in validity theory. The presumed desirable and undesirable consequence of various assessment practices have provided much of the support for newer performance assessment methods and assessment-oriented educational reform efforts. Facet 3 of Table 1 concerns the intended and unintended consequences of interpreting student performance, "the appraisal of value implications of score meaning, including value implications of the construct label itself, of the broader theory that undergirds construct meaning, and of the

still broader ideologies that give theories their purpose and meaning ." (Messick, 1995, p. 748). For example, does the way proficiency is interpreted have important consequences for particular groups of students? Facet 4 concerns the intended and unintended social consequences of using an assessment system. For example, did a given assessment practice have the desired effect of leading the students, teacher, and curriculum developers to focus more on higher-level understanding? Did completing the assessment lead them to integrate and organize what they already knew or did it lead them to doubt and question what they had learned?

Messick (1994, 1995) and others (e.g., Linn, Baker, & Dunbar, 1991) argue that it is particularly important to study the consequences of performance-based assessments because they promise positive consequences for learners and learning. These consequences are often cited to justify the added expense of performance-based assessment and to justify potential compromises to evidential validity. Messick further points out that it is wise to look for both the actual and *potential* consequences. Anticipation of likely outcomes helps us find side effects and determine what kinds of evidence are needed to monitor them. Such anticipation "may alert one to take timely steps to capitalize on positive effects and ameliorate or forestall negative effects" (Messick, 1995, p. 774).

Advancing a validity perspective that follows from the assumptions on learning and assessment embraced in our work (and outlined above), Frederiksen and Collins (1989) further emphasized the consequences of assessment practices by introducing the notion of *systemic validity*:

> A systemically valid test is one that induces in the educational system curricular and instructional changes that foster the development of the cognitive skills that the test is designed to measure. Evidence for systemic validity would be an improvement in those skills after the test has been in place within the educational system for a period of time (p. 27).

Frederiksen and Collins propose a set of principles for the design of systemically valid assessment systems, including the *components* of the system (a representative set of tasks, a definition of the primary traits for each subprocess, a library of exemplars, and a training system for scoring tests), *standards* for judging the assessments (directness, scope, reliability, and

transparency) and *methods for fostering self-improvement* (practice in self-assessment, repeated testing, performance feedback, and multiple levels of success)

In our opinion, Frederiksen and (1989) Collins advance a lofty, but worthy benchmark for evaluating assessment practice. Furthermore, documenting a measure of systemic validity will address the important concerns that have been raised by the assessment community regarding the potential negative consequences of short-answer paper and pencil assessment measures. Like many others, our assessment effort was constrained to such a format. We agree with Stiggins (1994) and others that, despite their limitations, paper and pencil assessments can be thoughtfully used to many some aspects of higher-order domain-specific understanding and to further develop that understanding.

In developing a framework for our validity inquiry, we found Moss' (1993) and Shepard's (1993, 1996) criticism of Messick's (1989) perspective invaluable. We initially struggled with the distinction between Messick's different facets with what Messick describes as the "progressive" nature of the framework, where construct validity appears in every cell, with something more added in each subsequent cell. Shepard (1993, p. 427) argues that Messick's faceted presentation implies that "values are distinct from a scientific evaluation of test score meaning" and "implicitly equates construct validity with a narrow definition of score meaning." Furthermore, the sequential segmentation of validity "gives researchers tacit permission to leave out the very issues which Messick has highlighted because the categories of use and consequences appear to be tacked on to 'scientific' validity which remains sequestered in the first cell." In our case, Messick would have us first evaluate the validity of our interpretation of student scores on our assessments (i.e., whether students perform poorly or well for reasons other than what we anticipated) before considering the consequences of our interpretation. Clearly though, the validity of our interpretation of performance is strongly impacted by the consequences of that interpretation. If students do not even try to finish a test because they are not being graded (i.e., minimal consequences of test use), then scores are an invalid depiction of student understanding *a priori*.

While Shepard agrees with Messick about the scope and range of validity inquiry, her differences with Messick's presentation have important implications for the way such inquiry is carried out. Shepard argues that Messick's framework does not help identify which validity questions are essential to support a test's use. This concern seems particularly relevant given

BEST COPY AVAILABLE

typically limited resources available for validity inquiry and the difficulty in prioritizing validity research questions. Indeed, we initially intended to focus only on "construct validity" because of the complexities of studying consequences of the assessments.

As an alternative to Messick's conceptualization, Shepard (1993, p. 428) equates construct validity "with the full set of demands implied by all four cells, which all involve score meaning." In light of the dilemma described above, Shepard insists that "intended effects entertained in the last cell are integrally part of test meaning in applied contexts." In order to provide "a more straightforward means to prioritize validity questions," Shepard suggests that

> validity evaluations be organized in response to the question "What does the testing practice claim to do?" Additional questions are implied: What are the arguments for and against the intended aims of the test? and What does the test do in the system other than what it claims, for good or bad? All of Messick's issues should be sorted through at once, with consequences as equal contenders alongside domain representativeness as candidates for what *must* be assessed in order to defend test use (1993, p. 429-430).

This view of validity inquiry draws strongly from Cronbach's (1988, 1989) concept of validation as evaluation argument, which in turn draws strongly from insights in program evaluation regarding the nature of evidence and argumentation, the posing of contending validity questions, and the responsibility to consider all of the potential audiences affected by a program. Cronbach has pointed out that program evaluators do not have the luxury of setting aside issues in the way that basic researchers typically do. Limited time and resources typically available to program evaluators forces them to identify the most relevant questions, and assign priorities depending on issues such as prior uncertainty, information yield, cost, and the importance of the questions for achieving consensus in the relevant audience.

Kane's (1992) extension of Cronbach's approach conceptualizes validation as the evaluation of interpretive argument.

> To *validate a test score interpretation* [including test uses] is to support the plausibility of the corresponding interpretive argument with appropriate evidence. The argument-based approach to validation adopts the interpretive argument as the framework for collecting

> and presenting validity evidence and seeks to provide convincing evidence for its inferences and assumptions, especially its most questionable assumptions. (1992, p. 527)

Drawing from literature on practical reasoning and evaluation argument, Kane identifies the criteria for interpretive argument as the following: (a) the argument must be clearly stated so that what is claimed is know; (b) the argument must be coherent in the sense that conclusions follow reasonably from the assumptions; and (c) assumptions should be plausible or supported by evidence, which includes investigating plausible counterarguements.

In summary, the interpretive argument approach described by Kane, along with Shepard's characterization of prioritizing one inquiry allow us to examine the validity of our assessment system in accordance with the assumptions on knowledge, transfer, and assessment described above. Following is a description of the assessment system we developed and the way we conducted this inquiry.

Method

Assessment System

The larger research agenda dictated several typical constraints for our assessment system. It needed to be paper-and-pencil, easy to score and interpret, and appropriate and fair for use in both implementation (i.e., GenScope) and comparison classrooms. Additionally, the assessment system needed to satisfy both formative and summative assessment goals, capture the full range of genetics reasoning within and between the various levels of biological organization, and be consistent with current understanding of the development of reasoning in introductory genetics (e.g., Stewart & Hafner, 1994; Kindfield, 1994).

Several design-implementation-revision cycles during the project's first year yielded two instruments. Both instruments were designed around fabricated species with simplified genomes consisting of three chromosomes and a handful of characteristics. The "NewWorm" was intended for younger and/or academically at-risk students, whereas "NewFly" was intended for older and/or college-bound students. As shown in the sample problems in Appendix A, the NewWorm provided some explicit genotype/phenotype relationships to scaffold the most basic understanding (i.e., the relationship is provided for the body-type characteristic but not for mouth type). On the NewFly assessment none of these relationships were provided for any of the

characteristics. While we used both of the assessments in the inquiry described here, we chose to use only the NewWorm in our subsequent investigations because it captures a broader range of expertise.

As shown in Table 2, we systematically varied the level of domain reasoning in our assessments along two dimensions. The type of reasoning assessed ranged from the simple cause-to-effect problems traditionally associated with secondary genetics instruction, to the more complex effect-to-cause problems that require the higher-level reasoning associated with domain expertise (and ostensibly afforded by GenScope). Both cause-to-effect and effect-to-cause reasoning were assessed in within-generation problems and in (more complex) between-generation problems. As shown by the vertical axis of Table 2, items within the various problem types ranged from the simple aspects of inheritance to the more complex aspects such as sex-linkage.

The cause-to-effect between-generation problems (the classic Mendelian inheritance problems that represent the typical extent of introductory genetics) varied on several additional dimensions. We included both categorical (*yes, maybe, no*) and more difficult proportional (*0, 1/8, 1/4, 1/2,* etc.) reasoning, and both monohybrid and (more difficult) dihybrid inheritance. Additionally, dihybrid inheritance included both unlinked and (more difficult) linked genes.

In keeping with the contemporary assessment perspectives outlined above, our assessments were designed to scaffold student problem solving across the increasingly difficult items. Specifically, we expected that solving the simpler initial problems would leave students with understanding (e.g., of the organism, our representational scheme, etc.) and self-confidence needed to solve the much more difficult problems later on.

Inquiry Framework

Our validity inquiry was organized around Messick's (1995) "six distinguishable aspects of construct validity". Table 3 provides a detailed description of the six and a list of the validity issues associated with each. Following the interpretive argument approach to validity inquiry advanced by Kane (1992) and Shepard (1993), we first defined the arguments that we anticipated making with or about our assessment system, and then exhaustively considered the potential threats to the validity of those arguments. Research priorities were established by weighing our concern over the particular threat with the resources needed to investigate it. The nature of our

inquiry ranged from incidental to explicit. The middle column of Table 3 summarizes inquiry methods for each aspect, and additional methodological details follow.

Readers should note that the segmented presentation does not imply the existence of six different types of validity. Like Messick's characterization, our inquiry reflects a unified concept of validity. As such, validity can neither rely on nor require any one form of evidence, and some forms of evidence may be forgone for other forms of evidence. "What is required is a compelling argument that the available evidence justifies the test interpretation and use" (Messick, 1995, p. 744). Our investigation bears out Messick's argument that the distinction between the six aspects provides "a means of addressing functional aspects of validity that help disentangle some of the complexities inherent in appraising the appropriateness, meaningfulness, and usefulness of score inferences" (1995, p. 744).

Content-related inquiry. The "content relevance, representativeness, and technical quality" (Messick, 1995, p. 745) of our assessment system was implicitly supported by having a nationally recognized content expert (the third author) lead the assessment team, and via routine feedback from teachers and content experts on the development/ implementation team. Once the assessments were developed, content was explicitly validated via review by outside content experts (both university-based science education researchers with extensive secondary biology teaching experience) and by comparing the assessments to the the biology content standards published by the National Research Council (1996).

In a somewhat novel aspect of our inquiry, we developed and validated a framework for documenting the degree of transfer represented by particular assessment performances. This framework was used to consider whether the curriculum activities and the classroom teaching practices corrupted the assessment activities (i.e., by reducing complex problem solving activities into simple algorithm or pattern recognition exercises). First we documented the number and nature of transformations between the assessment environment and the GenScope environment (and other likely comparison genetics learning environments). We further validated our assumptions about the learning environment by observing selected sessions in GenScope classrooms and interviewing the teacher. When paired with the results from the substantive inquiry (described below), this inquiry yielded a detailed framework for considering the degree of transfer represented by particular levels of assessment performance.

Structural, external, and generalizability inquiry. The "fidelity of the scoring structure to the structure of the construct domain", "extent to which scores' relationships with other measures and behavior reflect domain theory" and "extent to which score properties and interpretations generalize" was examined primarily via assessment scores from 13 high-school Biology classrooms before and after genetics instruction (including three classrooms where GenScope was implemented). These scores were analyzed using multi-faceted Rasch scaling (Linacre, 1989). This scaling method locates each assessment item and each individual's pretest or posttest performance on one linear scale. This yields an estimate of the relative "difficulty" of each item and the relative level of proficiency represented by each student's test performance, all using a common metric, along with data indicating the precision of the entire scale as well as each individual's and each item's fit on that scale.

While piloting the first version of the instrument, the item fits were used to flag potentially problematic items (e.g., items answered correctly by the less proficient students and/or incorrectly by the more proficient students), these items were examined and some were revised or removed. In the present inquiry, the scale scores for each item were used to validate our assumptions about domain structure, primarily by documenting whether increasingly more expert or more complex items were, in fact, more difficult. Similarly, the scale scores for each individual's pretest or posttest were used to validate assumptions about expected group differences and the effects of instruction. The reliability indices associated with the entire set of items and with the entire set of individuals informed assumptions about generalizability. Additionally, inter-rater reliabilities calculated by having multiple scorers score a subset of the assessments were used to validate the structural assumptions inherent in the scoring key.

Substantive inquiry. The "theoretical rationale for response consistency" was examined with a variety of methods. Before completing our assessment, students in the two GenScope pilot classrooms completed a "very near-transfer" GenScope quiz that assessed their ability to solve versions of selected assessment problems created using screen captures of the GenScope environment and the familiar GenScope dragons. Student performance on these items was examined in light of the GenScope curricular activities to determine whether students were actually learning the underlying domain concepts while completing those activities (our initial

observations suggested that many were not). Then, each student's performance on GenScope quiz items was examined in light of that student's performance on the corresponding NewFly items. It was expected that some (but not all) of the students who were able to solve a particular problem on the GenScope quiz would fail to solve the corresponding ("far transfer") problem on the NewFly quiz. Conversely, it was expected that few students would fail to solve a problem on the GenScope quiz but correctly solve the corresponding problem on the NewFly assessment.

When the NewFly posttest was administered in the GenScope pilot classroom, substantive validity was further investigated using videotaped think-alouds of four students solving the assessment problems and using videotaped interviewer probes of apparent understanding of assessed concepts in ten additional students. In the former, the first author provided an explanation of thinking aloud and a short practice session (following Ericsson and Simon, 1982) and then prompted students to continue thinking aloud as they progressed through their posttests. In the latter, students were videotaped while the interviewer went over already-completed posttests, probing the reasoning behind each response by gently challenging students on correct answers and providing hints and scaffolding for incorrect answers. Both procedures were used to help validate the accuracy of our interpretation of scores by looking for ways that students who seemed to understand the targeted concept failed to solve the corresponding problems (i.e., "construct irrelevant difficulty") and for ways that students got correct answers without the requisite understanding ("construct-irrelevant easiness"). The latter often occurs when critical aspects of test materials are well known to only some examinees, leading to invalidly high scores for those individuals. It was particularly important for us to look for construct-irrelevant easiness that might have been caused by the GenScope learning environment, as this would invalidate comparisons of learning between GenScope and comparison classrooms.

Consequential inquiry. The consequences of our assessment practice were considered relative to (a) the GenScope software and associated curriculum, (b) the learning environments in the three implementation classrooms, and (c) the students in those three classrooms. We first documented the changes in the curriculum and the software itself that could reasonably be attributed to the assessment development efforts and early assessment results. Then, in light of our assessment activity, we observed the learning environments, administered short surveys

alongside the assessments, interviewed two GenScope teachers, and interviewed five students in the GenScope implementation classrooms

Results

Reflecting our interpretive approach, our findings consist of warrants for the arguments we wished to make, rather than positivist "proofs" of validity

Content-Related Validity

Our analysis revealed that our assessments covered only a portion of the genetics content in the secondary school biology standards developed by the National Research Council (NRC, 1996) In light of the broad focus of the content standards, our assessments represented a narrower focus on reasoning about inheritance There were other aspects of genetics that were included in the GenScope curriculum (and many more that could have been included but were not) However, we elected to focus more specifically on what was emerging as the core of the GenScope curriculum and what is often the entire scope of secondary genetics instruction The two outside experts confirmed that our coverage of this aspect of the domain was very thorough, both in terms of the topics and the scope of reasoning around those topics While beyond the scope of the present research, our efforts highlight the tension between depth and breadth in curriculum and assessment practice and standards

The general consensus of the members of the assessment team and the larger GenScope team, along with the implementation teachers and the outside experts was that very few secondary school students ever develop the level of expertise represented by the most challenging problems on the assessment This was appropriate given the potential affordances of GenScope environment, our expectation that the design of the assessment would also scaffold student performance, and the need to capture the entire range of proficiency in our sample

Regarding transfer, our examination revealed the number and nature of transformations that GenScope students had to negotiate in order to succeed in the assessment environment These included *organism* (e g , GenScope dragons vs the NewWorm organism in our assessment), *traits* (e g , dragon's horns vs NewWorm's body shape), *representation* (e g , GenScope windows vs paper and pencil representations of those windows vs conventional genetics diagrams, text, etc), *genotypic configuration* (the *XY* females in most organisms and in

the NewFly assessment and the *XY* males in GenScope and the NewWorm assessment), *social context* (working with other students vs working individually), and *motivational context* (the typically ungraded GenScope activities vs graded assessment performance) We concluded that successful assessment performance represented a non-trivial transfer of understanding from the GenScope environment or other conceivable genetics learning environments

Structural Validity

The fit indices and scale scores derived from the Rasch scaling were examined to validate our assumptions about the development of expertise in the domain that we attempted to represent within and across the different types of problems Item fit indices show how well the relative difficulty of the various items was explained by the Rasch model Based on a standard normal curve, we would expect 95% of the items to fall within ± 2 0 SD, the number of items in excess of 5% outside of this range indicates the presence of variance that is not explained by the Rasch model On the NewFly, 40 of 56 items (71%) had standardized infit MSE within ± 2 0 SD (and 52 of 56 within ± 3 0 SD) Reflecting the fact that it was in essence a further refinement of the NewFly assessment, the fits for the NewWorm were better 50 of the 60 items (83%) had standardized infit MSE within ± 2 0 SD (and 57 of 60 items within ± 3 0 SD) The Rasch modeling also confirmed that our assessments captured a broad range of proficiency The separation index (a measure of the spread of the estimates relative to their precision) was over 5 0 for both instruments Loosely interpreted, this means that the precision of our assessments allow us to differentiate between five statistically significant intervals of proficiency in these populations This is supported by the fact that the reliability of the separation index for the items was 96 for NewWorm and 84 for NewFly, confirming that we had a wide range of item difficulty Similarly, the Rasch model revealed high reliabilities for students (79 for NewWorm, 87 for NewFly) indicating that these items were able to distinguish between student

Of primary interest in our analysis was the structure of the construct as revealed by the relative difficulties of the items within the assessments, in light of our assumptions about the development of domain expertise As described earlier (and shown on Table 1), we started with strong assumptions about the relative difficulty of the various items Figure 2 and Figure 3 show how the mean difficulties of the various clusters of NewWorm and NewFly items validated those assumptions First, we note that the item structure was generally replicated across the two

instruments. Across aspects of inheritance (i.e., from left to right), effect-to-cause reasoning was more difficult than cause-to-effect, and between-generation was more difficult than within-generation[1]. Across reasoning types (bottom to top), items involving complex aspects of inheritance (i.e., X-linkage) were more difficult than items involving simpler aspects.[2] Additionally dihybrid inheritance items involving linked alleles (thus requiring understanding of meiotic events to solve) were much more difficult than items that did not include genetic linkage.

Additional results not shown in Figure 2 and 3 further confirm our assumptions about the relative item difficulties. For the between-generation cause-to-effect problems (i.e., traditional Mendelian inheritance problems), items requiring probabilistic (e.g., 1/1, ½, ¼) reasoning were more difficult that items requiring categorical (yes, maybe, no) reasoning (+ 1.57 vs. −1.11 logits for NewWorm, + 1.15 vs. - .52 for NewFly). The item indices also confirmed that the items involving alleles for which we provided explicit genotype-phenotype relationships on the NewWorm (in order to scaffold very basic understanding) were easier than items that required the student to infer genotype-phenotype relationship (-2.42 vs. −1.47 logits for the cause-to-effect within-generation problems).

Substantive Validity.

Regarding our interpretation of assessment scores as evidence of understanding, our examination of students' answers on the NewFly assessment relative to their answers on the GenScope quiz revealed only a handful of cases where students failed to solve one of the "very near transfer" problems on the GenScope quiz yet provided a correct answer on the corresponding NewFly problem. Conversely, on each of the GenScope quiz items, only a subset of the students who solved a given problem went on to solve the corresponding problem on the NewFly assessment. This indicates both that our assessment minimized variance due to factors that we considered irrelevant to domain reasoning and that the assessment problems did require a reasonable transfer of understanding from the GenScope learning environment. Our

[1] Inadvertantly, the within-generation cause-to-effect items were not included this version of the assessment.
[2] An exception on both instruments was for the within-generation effect-to-cause problems, where problems involving X-linked genes were less difficult than problems involving autosomes. However, these are very simple problems that can be solved with little or no domain knowledge—essentially by identifying the appropriate phenotype (the expression of the trait, such as flat vs. round body) for a given allelic combination (e.g., BB, Bb or bb). In retrospect, it is not surprising that the difficulties for such items do not fully reflect our assumptions about domain reasoning.

BEST COPY AVAILABLE

observations of two classrooms where GenScope was implemented and examination of the existing curricular activities further validated our assumptions about the degree of transfer represented by the various assessment items. We determined that the few specific assessment items that might have been corrupted by particular kinds of instruction, were, in fact not corrupted[3]

The interviews and think-alouds generally revealed that students solved the NewFly problems the way we expected (except for one problem that was subsequently eliminated) With the exception of the most simple problems at the beginning of the assessment, there was little evidence of students "guessing" the correct answer. However, there were several of examples of students using the various cues included in the items (and in some cases using their answers to previous items) to figure out the correct answer to the more difficult problems. Given that these individuals could not be said to have initially "known" the answer, this might be characterized as "guessing" from a conventional assessment perspective, given that we designed the assessment to scaffold student problem solving, we view such instances as further validation of our assumptions about domain reasoning (e.g., that experts rely extensively on precisely such scaffolding to solve domain problems) and initial evidence of positive consequential validity This illustrates the paradox identified by Wiggins (1993) whereby the complexity of the assessment context is made manageable by the clues in that same context. We also found that even with extensive interviewer probing and scaffolding, students were generally not able to provide a correct answer (or demonstrate the target understanding) for items that were initially answered incorrectly. Thus we concluded that the complex context of the assessment successfully scaffolded domain reasoning without introducing construct-irrelevant easiness Given that construct-irrelevant variance has been identified as the major threat to the validity of this type of complex performance assessments (Messick, 1995), these are key findings in our inquiry

External Validity

These results concern the correspondence of students' assessment scores with other

[3] We were particularly concerned with the single-generation pedigree problems that ask whether a particular parental-offspring triad represented a dominant, recessive, or indeterminate mode of inheritance, and with the dihybrid inheritance problems involving linked alleles. Particular instructional treatment might have reduced these

 BEST COPY AVAILABLE

external indicators of proficiency. Figures 4 and 5 show the mean scaled student scores before and after instruction in the classrooms that used the NewWorm and the NewFly assessments, respectively. These figures show the change in mean proficiency (in logits) on the same scale as Figures 2 and 3, respectively, making it possible to consider mean level of proficiency in each classroom in terms of the level of domain reasoning.[4]

The first consideration is whether expected group differences are observed, regardless of instruction. Figure 4 shows that the 10th graders who studied genetics in Biology 1 (college track) were more proficient than their school mates in Biology 2 (general track), who in turn were more proficient than the disadvantaged inner city students whose genetics instruction took place within a general sciences course[5]. Similarly Figure 5 shows that the overall mean proficiency in the three Bio 1 classrooms was higher than in the three Bio 2 classrooms.[6]

In terms of the impact of instruction, all of the classrooms showed gains in reasoning ability from the pretest to the posttest. Overall proficiency in the three classrooms that completed the NewWorm assessment before and after instruction increased from .19 logits to .79 logits.[7] Figure 4 shows that increases in mean proficiency in domain reasoning ranged from 0.5 to 0.75 logits of the roughly 4.5 logits represented by the different item clusters. Such gains are quite modest in light of the roughly 4.5 logit range represented by the various item clusters shown on Figure 2. The genetics curriculum in the suburban comparison classrooms was fairly conventional text-based instruction that combined teacher-led classroom activities and a self-paced "drill and practice" workbook. The curriculum in the urban GenScope pilot classroom consisted largely of curricular activities designed around the GenScope software, while the curriculum in the urban comparison classroom consisted of conventional textbook-based instruction and activities (all three urban classrooms were taught by the same teacher). While the gains were somewhat larger for the two the suburban comparison classrooms than in the inner-

---

problems to a simple pattern recognition task, allowing students to solve them without the relevant domain understanding.

[4] Because of the differences in the way the two assessments presented information, the data from the two assessments cannot be scaled together.

[5] Mean of pretest and postest scores for students that completed both were .93, .22, and -.13 logits, respectively, for the three groups, $F(2,42) = 4.93$, $p = .012$.

[6] Mean of pretest and posttest scores for student that completed both were -.35 and -1.09, respectively, for the the two groups, $F(1,102) = 17.6$, $p < .0001$.

[7] $F(1,41) = 13.6$, $p = .001$

city GenScope pilot classroom, the difference was not statistically significant [8] Encouragingly, the mean posttest score in the GenScope pilot classroom was still higher than in the other two urban classrooms [9] It should be noted that the students in all three of the urban classrooms were 9th graders who were among the most academically at risk students within a generally disadvantaged population

In the six classrooms that completed the NewFly assessment, mean proficiency increased from - 64 logits to .27 logits [10] Figure 5 shows increases ranging from 0 5 to 1 3 logits of the roughly 3 5 logits represented by the different NewFly item clusters shown in Figure 3, with gains in the two Bio 2 classrooms participating in the GenScope pilot implementation nearly identical to the gains in the comparison Bio 2 classroom at the same school where a mix of teacher-directed instruction and a self-paced workbook was used to teach introductory genetics Thus, we conclude that students in the classrooms that participated in the pilot implementation of GenScope made the same modest gains in domain reasoning as students in classroom using more conventional curricula In terms of our operationalization of domain reasoning, this gain is roughly akin to the difference between Mendelian inheritance problems involving basic autosomal traits and problems involving X-linked traits Such disappointingly results are consistent with the prior research (as reviewed by Stewart and Hafner, 1994) As described below, this GenScope pilot implementation revealed many aspects of the software and the curriculum that needed further development While it is encouraging that students in the GenScope classrooms did as well as students in the comparison classrooms, this learning environment is expected to ultimately support much larger gains in reasoning than conventional curricula

Further evidence of external validity was provided by the fits for the individual student's scores at both pretest and posttest The Rasch person scaling results revealed very good fits for both instruments; Of the 143 students who took the NewWorm, 134 (94%) had a standardized infit MSE within ± 2 0 SD; Of the 243 student who took the NewFly, 232 (95%) had a

---

[8] $|F(2,41) < 1|$ The standard deviations for each class's pretest or posttest ranged from .55 to 1 1, except for the *GenScope* class at pretest (SD = 1 4) and the Bio 1 class at posttest (SD = .37)

[9] While students were not pretested in these other two inner-city classrooms, the teacher indicated that students in those other two classrooms were generally more proficient than the students in the classroom that piloted the *GenScope* curricula

[10] $F(1,102) = 149\ 8$, $p > .0001$ The standard deviations for each class's pretest or posttest ranged from .66 to .87, except for one *GenScope* class at pretest (SD = 1 5)

BEST COPY AVAILABLE

standardized infit MSE within ± 2 0 SD Both findings are consistent with our expectation based on a standard normal curve. This shows that the pattern of proficiency within and across individual tests was well explained by the Rasch model In practical terms, this indicates that students generally provided correct answers for increasingly difficult items only up to the extent of their proficiency, and that they did not miss a lot of easy items while getting difficult items correct

Generalizability

Reflecting our research priorities, generalizability inquiry was limited to the generalizability of scorers across raters This evidence was provided by conventional rater agreement indices Operational scoring for all of the assessments included in the present research was conducted by the same individual, who had completed prior undergraduate-level biology coursework. The majority of the items on both instruments requested short answer responses for which no scorer interpretation was needed Informal training on the use of the scoring key was provided when the assessments were piloted, so no formal training was needed to score tests in the present sample

Rater agreement was calculated for the 14 NewWorm items for which scoring required some interpretation The third author's scores on these items were then compared to those of the primary scorer Eight of these items that asked students to explain/justify their short answers on effect-to-cause reasoning problems (e g., "What is it about the offspring data that indicates whether the gene is autosomal or sex linked?") and were scored on a three point (none/partial/full) scale Inter-rater agreements (Pearson's r) on these eight items averaged 93 and ranged from 81 to 1 0 Cohen's κ on the eight items averaged 86 and ranged from 69 to 1 0 On the six remaining problems requiring interpretation, inter-rater agreement averaged only 62 and ranged from 0 to .93. However, these items proved exceedingly difficult to answer with roughly two-thirds of the students leaving them blank and as few as one or two students receiving full credit As such, these items were completely revised following this implementation While these numbers indicated a need for further development of the scoring keys, they did show generalizability of student scores across multiple raters

Consequential Validity

Regarding the intended positive consequences of our assessment practice, the results were decidedly mixed. While there were some positive changes to the software and the curricular activities made in response to the assessment team's input and disappointing early pilot results, many of the indicated changes were put off until after the implementation cycle. Perhaps most importantly, there were no curricular activities used in any of the GenScope pilot classrooms that explicitly targeted between-generation effect-to-cause reasoning in the GenScope curricula used in the three classrooms. Hence, the finding that only a handful of students in any of the GenScope classrooms demonstrated that level of posttest proficiency was not surprising.

The consequences of our assessment practice, positive or negative, for the learning environment and the learners were quite limited. While the think-alouds conducted as part of the substantive inquiry certainly suggested that students were constructing useful knowledge in the process of completing the assessment, additional research is needed to verify the nature and extent of that learning. The students in the GenScope classrooms were certainly aware that they were participating in an activity that was important to their teacher and the outside researchers who were visiting their classroom. However, student' comments on an open-ended survey administered with the assessment and an informal interviews with students and the teacher in the two GenScope classrooms that used the NewFly assessment revealed that the students felt that the assessment was not very connected to curriculum—in their words, not "fair." While the students tried hard and wanted to do well (both for the grade assigned by their teacher and for the sense of accomplishment), it was clear that the students did not get the feedback needed to use the assessment system to help them learn.

Given that the involvement of the assessment team led to demonstrable desirable changes to the software and curriculum, our particular assessment practice seems to have avoided the most common negative consequence of assessment practices—focusing the curriculum on the sorts of basic factual knowledge that can be readily assessed in a multiple-choice format. The only possible negative consequence that the students' generated in our informal interview was the time spent on the assessment could have been spent doing GenScope activities. However, many of the students agreed that they had actually learned something through completing the assessment, despite the lack of feedback. When queried further, several students specifically

BEST COPY AVAILABLE

volunteered having had "a-ha" experiences where they suddenly came to understand something that they had been confused about, while completing the assessment.

Conclusions and Implications

The results strongly support the *evidential* validity of our assessment practice. This includes both our interpretation of assessment scores as evidence of domain understanding ("the evidential basis of test interpretation") and our use of these scores to evaluate learning in GenScope environments ("the evidential basis of test use"). Together, these findings support our conclusion that GenScope and non-GenScope students make similarly modest gains in their ability to reason in the domain of introductory genetics.

In contrast, the results provide mixed support for the consequential validity of our assessment practice ("the consequential basis of test use and interpretation"). While our input as observers and outside evaluators had some positive consequences for the GenScope software and curricula, there was little evidence that our assessment practice helped students—compared to not using the system at all. Clearly our students were motivated to do well on the assessment system and there was some indication that students learned while completing the assessment. However, our assessment practice did not significantly, in the words of Frederiksen and Collins (1989) "foster the development of the cognitive skills it was designed to measure." In other words, our assessment system had yet to achieve *systemic validity*. Indeed, we failed to establish a complete set of what Frederiksen and Collins (1989) advanced as critical attributes of systemically valid assessments. Our system did include the necessary *components* (i.e., a representative set of tasks, a definition of the primary traits for each subprocess, a library of exemplars, and a training system for scoring tests) and *standards* for judging the assessments (directness, scope, reliability, and transparency). However, we believe that all of these components could have been better utilized to support student learning. More critically though, with, with the exception of supporting multiple levels of success, we failed to establish *methods for fostering self-improvement* such as practice in self-assessment, repeated testing, and performance feedback.

In order to enhance systemic validity in subsequent implementation classrooms (and to therefore help foster the higher-level understanding that we believe GenScope affords), we have developed a set of curriculum activities that use GenScope dragons to scaffold the domain

reasoning represented by the NewWorm and NewFly assessments. Initially, these so called "Dragon Investigations" evolved from the near-transfer items on the GenScope quiz. We examined the types of reasoning represented by the various assessment items and created worksheets that have students solve those same kinds of problems using the more familiar GenScope dragons. The activities were designed to be useful away from the computer, either as homework or in class. We also provided the teachers with answer keys for each worksheet that included detailed explanations of the relevant domain content in the context of solving the problem. (Appendix B contains an example of one of the student worksheet and the teacher version). We believe that along with the further refinement of the GenScope software and curricular activities, students in subsequent implementations will demonstrate dramatically larger gains in domain understanding.

More generally, we conclude that our investigation illustrates how recent advances in measurement can assist those concerned primarily with designing and implementing learning environments. Rasch modeling provided a wealth of useful analyses, and the new validity frameworks and methods provide a way to integrate the goals of two potentially antagonistic perspectives. For example, the planned changes to the GenScope curriculum may well compromise the evidential validity of our assessments in future evaluations. However, because we anticipate this conflict and have a means of interpreting the degree of compromise (by documenting the degree of transfer), the tradeoff between evidential and systemic validity can be thoughtfully considered. This makes it possible to maximize learning while still conducting rigorous evaluation of the environment.

References

Cronbach, L. J. (1988). Five perspectives on validity argumentation. In H. Wainer and H. Braun (Eds.), *Test validity*, (pp. 3-17). Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. E. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 147-171). Urbana, University of Illinois Press.

Ericsson, K. A., & Simon, H. A. (1982). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18* (9), 27-32.

Greeno, J. G., Collins, A. M., & Resnick, L. (1996). Cognition and learning. In D. Berliner and R. Calfee (Eds.) *Handbook of Educational Psychology*, (pp. 15-46). New York: MacMillan.

Horwitz, P., Neumann, E., & Schwartz, J. (1996). Teaching science at multiple levels: The GenScope program. *Communications of the ACM, 39*, (8), 127-131.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535.

Kindfield, A. C. H. (1994). Understanding a basic biological process: Expert and novice models of meiosis. *Science Education* 78, 255-283.

Linacre, J. M. (1989) *Many-faceted Rasch measurement*. Chicago, IL: Mesa Press.

Linn, R. L., Baker, E. L., & Dunbar, S. B., (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20 (8)*, 15-21.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23 (2)*, 13-23.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*, 229-258.

National Research Council (1996). *National science education standards*. Washington DC: Author.

President's Committee of Advisors on Science and Technology, Panel on Educational Technology (PCAST) (March, 1977). *Report to the president on the use of technology to strengthen K-12 education in the United States*. Author.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*, 404-450.

Shepard, L. A. (1996, April). *The centrality of test use and consequences for test validity*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Stewart, J., & Hafner, R. (1994). Research on problem solving: Genetics. In D. Gabel (Ed.) *Handbook of research on science teaching and learning* (pp. 284-300). New York: Macmillan.

Stiggins, R. J. (1994). *Student-centered classroom assessment*. Upper Saddle River, NJ: Prentice-Hall.

Vygotsky, L. S. (1978). *Mind in society: The development of higher mental processes*. Cambridge, MA: Harvard University Press.

Wiggins, G. (1993). Assessment: Authenticity, context, & validity. *Phi Delta Kappan, 75*, 200-214.

Wolfe, D., Bixby, J., Glenn, J., and Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. *Review of Research in Education, 17*, 31-74.

Table 1 Four facets of validity, after Messick, (1989, 1995)

	Interpreting Tests	Using Tests
Evidential Basis of Validity	1. The Evidential Basis of Test Interpretation (Is the assessment a valid measure of domain understanding?) Construct Validity	2. The Evidential Basis of Test Use (Is our assessment system a valid measure of learning in this particular environment?) Construct Validity + Relevance/Utility
Consequential Basis of Validity	3. The Consequential Basis of Test Interpretation (What are the consequence of our interpretation of performance for particular individuals or groups?) Construct Validity + Value Implications	4. The Consequential Basis of Test Use (What are the intended and unintended consequences of our assessment practice?) Construct Validity + Relevance/Utility + Value Implications + Social Consequences

Table 2 Dimensions of domain reasoning in *NewWorm* assessment and corresponding items

Aspects of Inheritance (Less Complex ↔ More Complex)	Domain Reasoning Types (Novice ↔ Expert)				
	Cause-to-Effect Within Generation	Effect-to-Cause Within Generations	Cause-to-Effect Across Generations	Effect-to-Cause Across Generations	Reasoning about the Meiotic Process
X-Linked/ Simple Dominance	Genotype-Phenotype Mapping 5.	Phenotype-Genotype Mapping 6	Monohybrid Inheritance II. 2	Monohybrid Inheritance III Eyelids Pedigree II Night Vision	Meiosis The Process
Autosomal/ Incomplete Dominance	Genotype-Phenotype Mapping 3. 7	Phenotype-Genotype Mapping 3. 5	Monohybrid Inheritance II. 1		Meiosis Gametes
Autosomal/ Simple Dominance	Genotype-Phenotype Mapping 2. 4. 6	Phenotype-Genotype Mapping 1. 2	Monohybrid Inheritance I	Monohybrid Inheritance III Texture Pedigree I Dominance Relationships Pedigree II Color Vision	
Sex Determin-ation	Genotype-Phenotype Mapping 1				

32

33

Table 3 Six Aspects of validity and associated inquiry methods and findings After Messick, (1995)

Aspect of Validity and Associated Validity Issues	Validity Inquiry Conducted (or Implied)	Validity Findings and/or Conclusions
CONTENT **"Content relevance, representativeness, and technical quality."** • Looking at the right things in right balance? • Anything important been left out? ("construct under-representation") • What are construct-relevant sources of task difficulty? • Appropriate sampling of domain processes?	1. Content expertise was present throughout assessment development 2. Requested independent outside domain experts to review our assessments 3 Analyzed correspondence between NSTA Genetics content standards and our assessments 4 Analyzed assessment tasks in light of the learning environment	1 A Kindfield is an acknowledged authority in development of expertise in the domain of genetics Domain experts on curriculum teams also reviewed content 2 Outside domain experts validated content structure 3 Assessments covered half of the standards 4 Analysis of the number and nature of transformations between learning environment and assessment environment documented the degree of transfer represented by the assessments
STRUCTURAL **"Fidelity of the scoring structure to the structure of the construct domain"** • Is the scoring structure (including task-weighting, partial credit, conditional scoring, etc) consistent with the domain structure?	1. Assessed inter-rater reliability 2. Used Rasch model to validate partial credit schemes 3 Used Rasch model to analyze item difficulties in in light of assumptions about domain reasoning 4. Compared item difficulties across NewFly and NewWorm	1 Inter-rater reliabilities were very high 2 Probability curves indicated that some partial credit items should be dichotomous 3 Item analyses revealed expected order of item difficulties effect-to-cause easier than cause-to-effect, simple modes of inheritance easier than complex modes 4 Item difficulty replicated across the two
SUBSTANTIVE **"Theoretical rationale for response consistency, including process models of task performance, along with empirical evidence that the theoretical processes are engaged by respondents.** • Are ostensibly sampled processes the ones that students actually engage in? • Are there plausible counter-interpretations for successful or poor performance?	1 Conducted think-alouds and interviews with 15 students during and after assessments to document solution processes and search for alternative interpretations 2 Compared videos of assessment solution processes with videos of students in classroom solving associated problems 3 Compared performance on "very-near transfer" quiz items and assessment items	1 Results used to revise or eliminate problematic items Sudents solved remaining problems the way that we expected (could not solve problems without relevant understanding, and could not guess difficult answers) 2 Analyses verified that learning environment didn't reduce complex problems into algorhythmic activities 3 Students didn't fail quiz item but solve corresponding assessment item (but did vice versa) Demonstrates appropriate transfer demands and argues against construct-irrelevant-variance

BEST COPY AVAILABLE

Table 3 Aspects of validity, inquiry methods, and findings (continued from previous page)

Aspect of Validity and Associated Validity Issues	Validity Inquiry Conducted (or Implied)	Validity Findings and/or Conclusions
GENERALIZABILITY **"Extent to which score properties and interpretations generalize across groups populations, settings and tasks."** • Is the interpretation of scores (as evidence of learning) reliable and does that interpretation generalize across contexts and populations?	1 Examined reliabilities in Rasch model 2 Documented inter-rater agreement	1 High reliabilities for individuals and items in Rasch model 2 Acceptable inter-rater agreement
EXTERNAL **"The extent to which the scores' relationship with other measures/ behavior reflect the expected relations implicit in domain theory"** • Does scores converge and diverge with other scores and behaviors sensibly? • Are the value implications of the score interpretation empirically grounded?	1 Compared scale scores before and after Genetics instruction 2 Compared classroom scale means for different groups of students	1 Scale scores were significantly higher following instruction 2 Means for different classrooms following instruction (AP, Bio 1, Bio 2) revealed expected group differences
CONSEQUENTIAL **"Intended and unintended consequences of interpreting and using the score"** • Are the short-term and long-term consequences of score interpretation and use supportive of the general testing aims? • Are there adverse side-effects?	1 Analyzed meeting notes and project emails regarding influence of assessment team on software and curricula development effort 2 Analyze software & curriculum in light of of assessment development process and assessment results 3 Observed classroom learning environment in light of assessment system 3 Adminstered surveys to students after they completed assessments 4 Conducted think-alouds with four students as they completed their posttest 5 Interviewed 15 students following assessments	1 Documented systemic influence of assessment effort 2 Numerous changes to software and curricula attributable to assessment development process and initial results 3 observations showed that assessments had modest positive influence on learning environment (i e , some clarification of focus) but little influence overall 4 Survey results indicated that students considered the assessments to be important and that they tried hard 5 Some evidence that students constructed useful understanding while completing assessments 6 Inteviews revealed that assessments were a strong incentive, but were not well linked to learning environments. Poor performance left some students demoralized

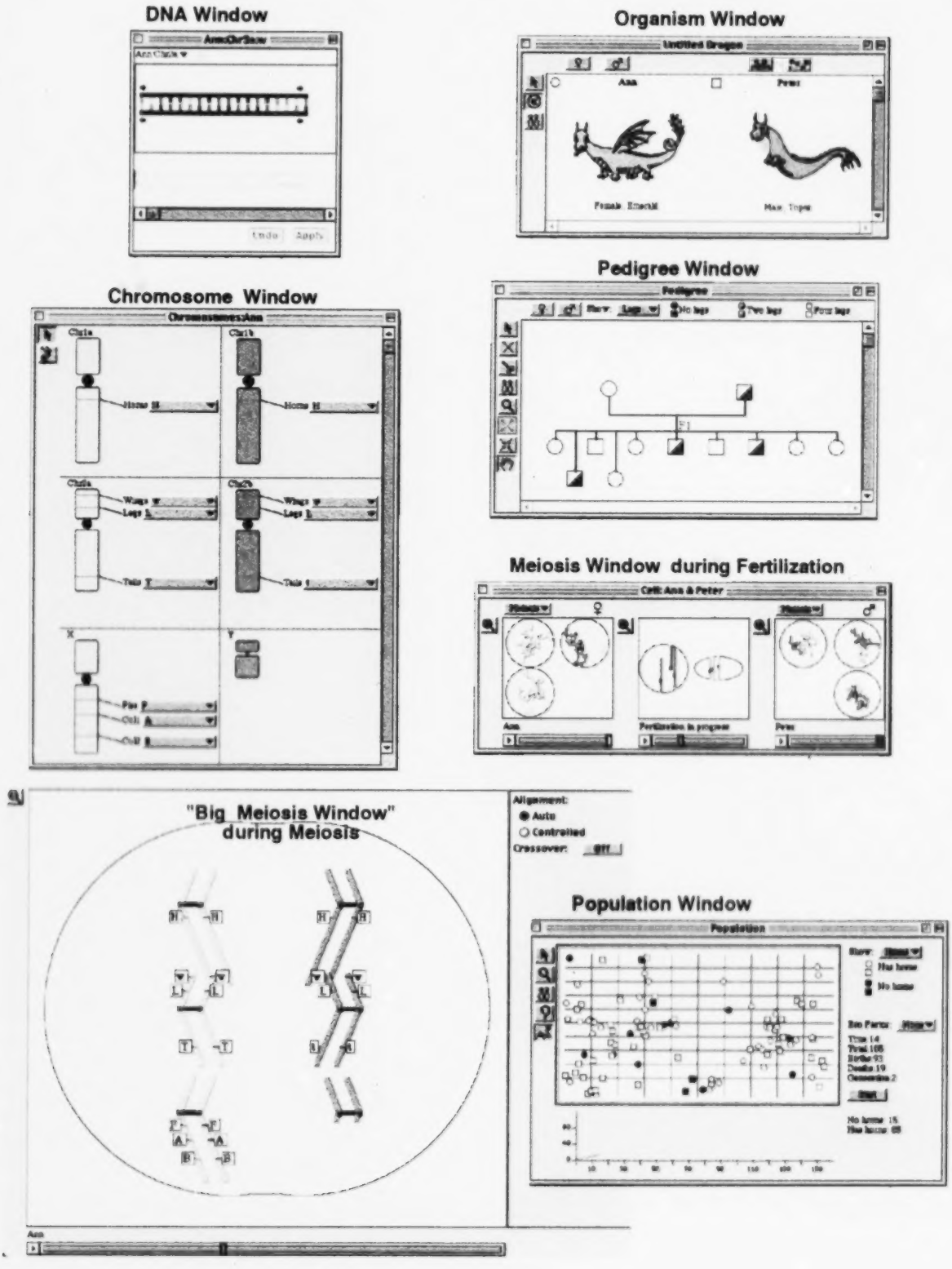
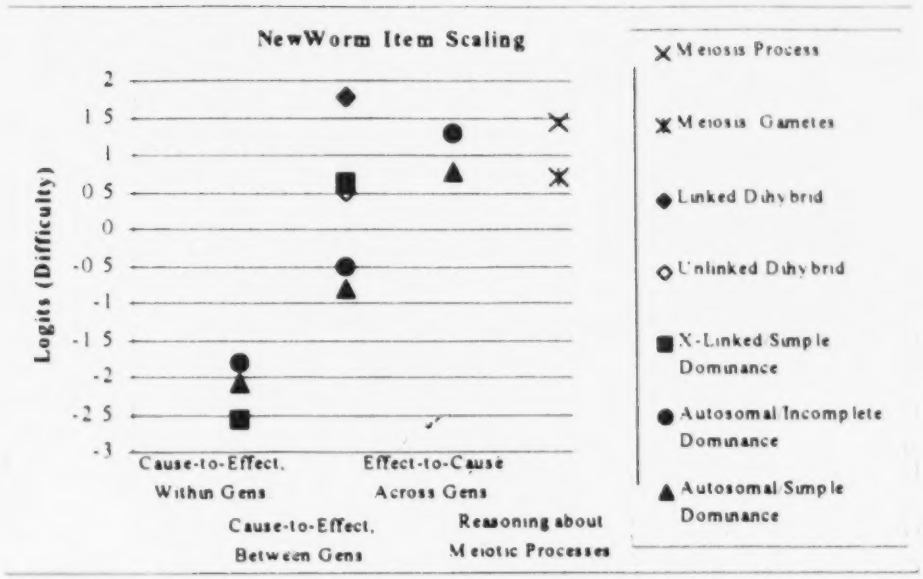


Figure 1. Examples of Screens in the *GenScope* Software.

BEST COPY AVAILABLE



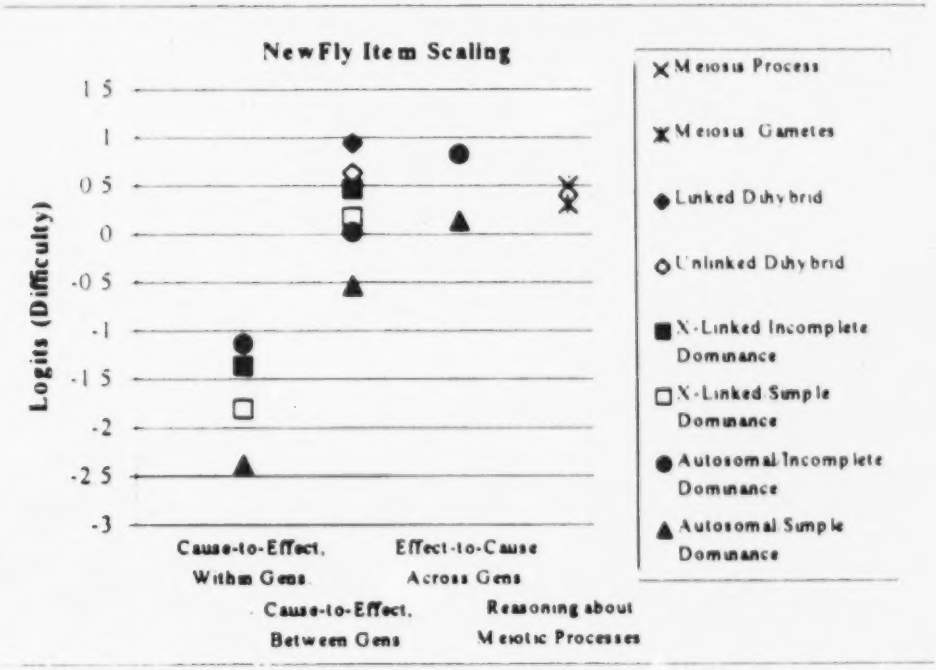
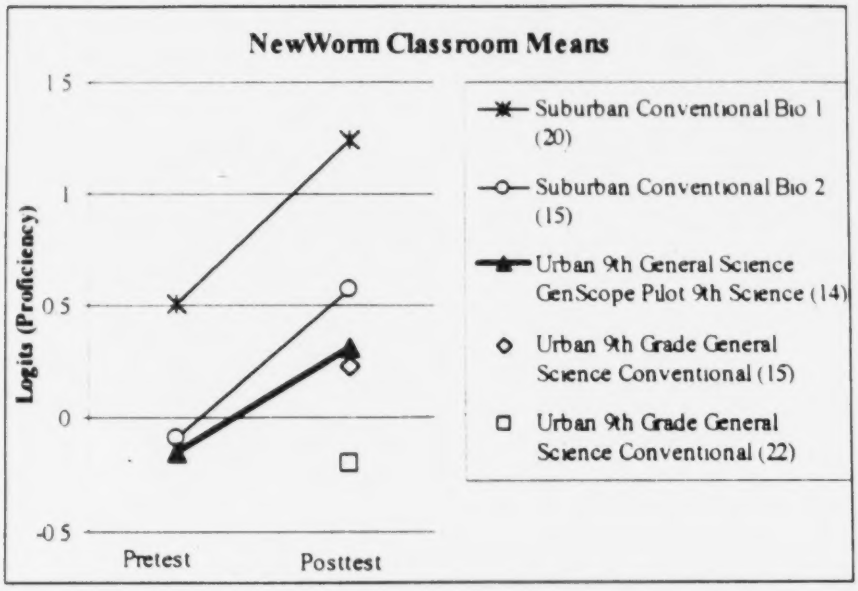


Figure 2 Item Cluster Difficulties for NewWorm Assessment Items

Figure 3 Item Cluster Difficulties for NewFly Assessment Items



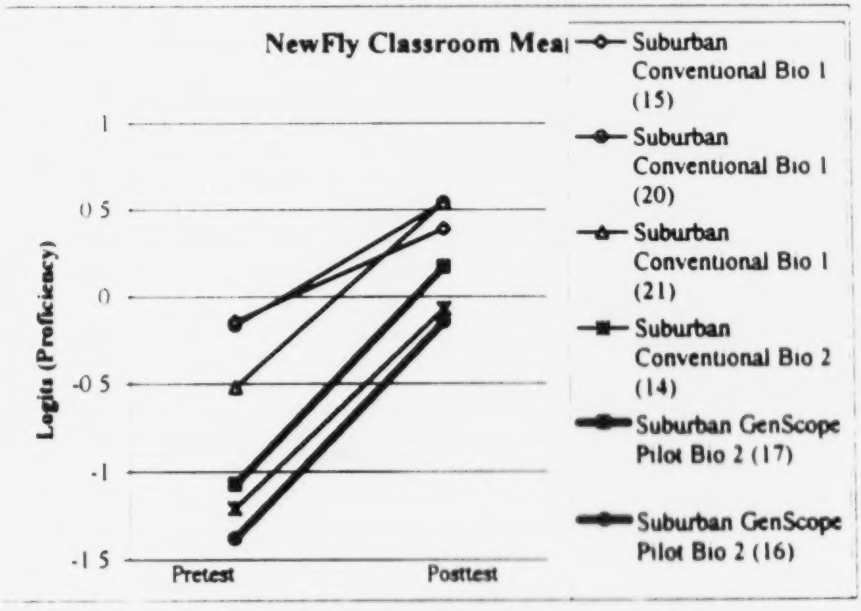


Figure 4 Classroom Mean Proficiency Before and After Instruction (NewWorm Classrooms)

Figure 5 Classroom Mean Proficiency Before and After Instruction (NewFly Classrooms)

BEST COPY AVAILABLE

Appendix A

Example Items from *NewWorm* Assessment

(with type of reasoning, aspects of inheritance, and item type indicated)

The NewWorm©

Copyright 1998, A. Kindfield & D. Hickey

Original image copyright 1997, William Wadsworth
Used with permission

The left box shows what we know about NewWorms' genes. The right box shows the genetic makeup of two NewWorms. Use this information to solve the problems below.

NewWorm Genetics

Body: Flat: **BB** or **Bb** Round: **bb**

Mouth: Oval: **??** Slit: **??**

Head: Broad: **??** Medium: **??** Narrow: **??**

Rings: No Rings: **RR** or **Rr** Rings: **rr**

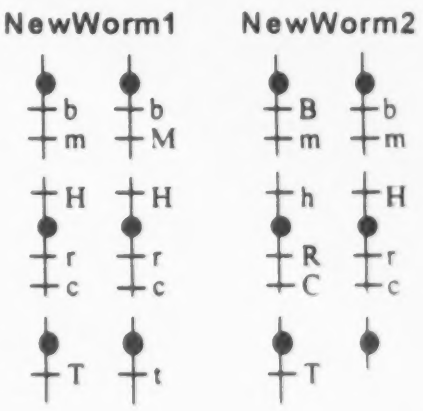
Color: Green: **CC** Brown: **Cc** Black: **cc**

Tail (Male): Pointed: **TT** or **Tt** Blunt: **tt**

Tail (Female): Pointed: **T–** Blunt: **t–**
(The Tail gene is on the **X** chromosome.)
(The **–** [dash] stands for the **Y** chromosome.)

Sex: Males: **XX** Females: **XY**

Two NewWorm Genotypes



GENOTYPE-PHENOTYPE MAPPING
(cause-to-effect, within generation)

Determine phenotypes (traits) from NewWorm1 and NewWorm2's genotypes:

	NewWorm1	NewWorm2
What body shape? (autosomal simple dominance)	1a.	1b.
What kind of tail? (X-linked simple dominance)	4a.	4b.
Male or female? (sex determination)	5a.	5b.

If the allele for **oval mouth (M)** is dominant to the allele for **slit mouth (m)**

What kind of mouth? (autosomal simple dominance with implicit genotype-phenotype relationship)	6a.	6b.

BEST COPY AVAILABLE

PHENOTYPE-GENOTYPE MAPPING
(effect-to-cause, within generation)

NewWorm Genetics	Two NewWorm Phenotypes	
	NewWorm3	NewWorm4
Body: Flat: **BB** or **Bb** Round: **bb**	flat body	round body
Mouth: Oval: **??** Slit: **??**	slit mouth	oval mouth
Head: Broad: **??** Medium: **??** Narrow: **??**	narrow head	medium head
Rings No Rings: **RR** or **Rr** Rings: **rr**	rings	no rings
Color: Green: **CC** Brown: **Cc** Black: **cc**	brown	green
Tail (Male): Pointed: **TT** or **Tt** Blunt: **tt**	blunt	pointed
Tail (Female): Pointed: **T–** Blunt: **t–** (The Tail gene is on the **X** chromosome) (The **–** [dash] stands for the **Y** chromosome)	male	female
Sex: Males: **XX** Females: **XY**		

For each characteristic, circle ALL of NewWorm3's possible genotypes.

Characteristic	NewWorm3					
1. Body	BB	Bb	bb	B–	b–	(Autosomal simple dominance)
2. Mouth	MM	Mm	mm	M–	m–	(Autosomal simple dominance)
3. Head	HH	Hh	hh	H–	h–	(Autosomal incomplete dominance)
4. Rings	RR	Rr	rr	R–	r–	(Autosomal simple dominance)
5. Color	CC	Cc	cc	C–	c–	(Autosomal incomplete dominance)
6. Tail	TT	Tt	tt	T–	t–	(X-linked simple dominance)

Remember

- the allele for oval mouth (**M**) is dominant to the allele for slit mouth (**m**) and
- the allele for broad head (**H**) is incompletely dominant to the allele for narrow head (**h**) and medium head is in between broad and narrow

MONOHYBRID INHERITANCE I
(cause-to-effect, across generations)

Figure out whether a baby produced by NewWorm1 and NewWorm2 will have a round body:

Color (autosomal incomplete dominance)

2a Will a baby be **brown?**

Definitely yes (categorical reasoning) Maybe Definitely no

2b What are the chances that a baby will be **green**?

0 1/4 1/2 3/4 1/1
(probabalistic reasoning)

DIHYBRID INHERITANCE
(cause-to-effect, between generations)

Use the NewWorm1 and NewWorm2 genotypes to answer these questions about their babies.

Color and Rings (autosomal incomplete and simple dominance, linked dihybrid)

2a Will a baby have a **brown body** AND **rings?** (categorical reasoning)

Definitely yes Maybe Definitely no

2b What are the chances that a baby will have a **black body** AND **rings?**
(probabalistic reasoning)

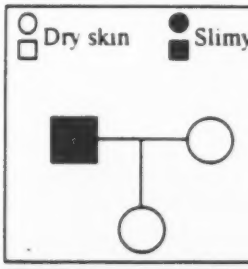
0 __ 1/8 __ 1/4 __ 3/8 __ 1/2 __ 3/4 __ 1/1 __

OR **impossible to tell from what's given** __

PEDIGREE I: DOMINANCE RELATIONSHIPS
(effect-to-cause, across generations)
(simple dominance–focus on dominance relationships)

Consider four other NewWorm characteristics–Skin, Nostrils, Eyes, and Tongue.

- Each characteristic has two phenotypes as shown with the pedigree.
- Females are represented by circles and males are represented by squares.
- Decide what each pedigree says about the dominance relationship between each pair of phenotypes.

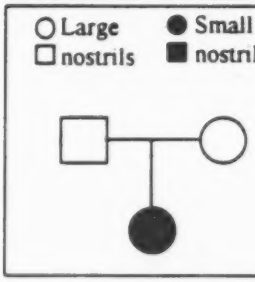


1. Having slimy skin is:

_____ definitely dominant

_____ definitely recessive

_____ impossible to tell from what's given



2. Having small nostrils is:

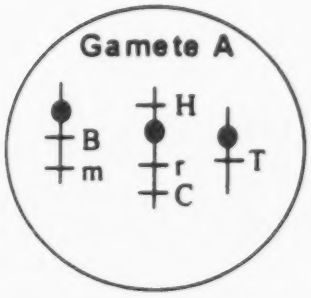
_____ definitely dominant

_____ definitely recessive

_____ impossible to tell from what's given

MEIOSIS: GAMETE A

(reasoning about meiotic processes)



1. Was crossing over necessary for NewWorm2 to produce Gamete A?

   Answer

1a. If you answered yes, circle the chromosome(s) in Gamete A that resulted from crossing over.

If you answered no, check here.

If you did not answer, do nothing.

MONOHYBRID INHERITANCE II: EYELIDS
(effect-to-cause, across generations)
(X-linked simple dominance)

Another inherited characteristic in the NewWorm is Eyelids. Both NewWorm1 and NewWorm2 have clear eyelids. However when you mate them and produce 100 offspring, you find:

- 74 (51 males and 23 females) have clear eyelids
- 26 (0 males and 26 females) have cloudy eyelids

Remember: Males are XX and females are XY

1. There are two alleles for Eyelids. Is the relationship between the two alleles simple dominance or incomplete dominance?

   Answer:

   1a. What is it about the **offspring** that indicates simple or incomplete dominance?

2. **If** one of the Eyelids alleles is dominant, which one is it (clear, cloudy, OR neither)?

   Answer:

   2a. What is it about the **offspring data** that shows you which, if any, allele is dominant?

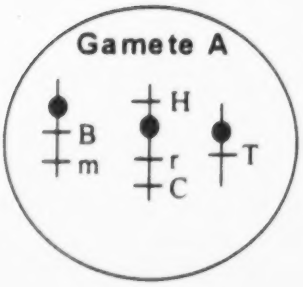
3. Is the gene for Eyelids autosomal or X-linked?

   Answer:

   3a. What is it about the **offspring data** that indicates whether the gene is autosomal or X-linked?

MEIOSIS: GAMETE A

(reasoning about meiotic processes)



1 Was crossing over necessary for NewWorm2 to produce Gamete A?

Answer

1a. **If** you answered **yes**, circle the chromosome(s) in Gamete A that resulted from crossing over.

If you answered **no**, check here

If you did not answer, do nothing.

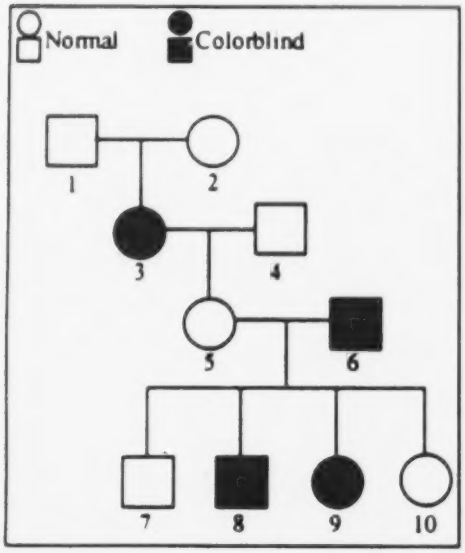
PEDIGREE II: COLOR VISION–AUTOSOMAL OR X-LINKED?

(effect-to-cause, across generations)

(autosomal simple dominance)

Consider another NewWorm characteristic–Color Vision.

- Color Vision has two phenotypes as shown with the pedigree.
- Females are represented by circles and males are represented by squares.
- **Remember**: Males are XX and females are XY.
- Decide if the pedigree is consistent with Color Vision being autosomal or X-linked.



1 Does the Color Vision gene **appear** to be autosomal or X-linked?

Answer

1a Using words and/or diagrams, explain your answer (use the numbers below each circle or square to refer to particular individuals)

2 Does this pedigree rule out the type of inheritance you did *not* choose?

Answer

2a. Using words and/or diagrams, explain your answer (use the numbers below each circle or square to refer to particular individuals)

BEST COPY AVAILABLE

Appendix B

Sample "Dragon Investigation"

(Student Worksheet)

From Offspring to Mode of Inheritance

We often don't know the genotypes of individuals or the genetics of the species for a particular characteristic. One way to figure out the genetics of a particular characteristic is to carefully study of the patterns of inheritance of phenotypes.

Fangs

Another inherited characteristic in dragons is Fangs. Both Sandy and Pat have no fangs. But when you look at 100 of their offspring, you find the following:

- 29 (13 males and 16 females) have fangs
- 71 (37 males and 34 females) have no fangs

Monohybrid Inheritance III: Phenotypes to Genotypes

Use the information about the offspring to explain the mode of inheritance. Remember that in dragons, males are XX and females are XY.

1. The Fangs gene has two alleles–*fangs* and *no fangs*. The relationship between the two alleles is **simple dominance** (rather than incomplete dominance).

   What is it about the **offspring phenotypes** that indicates that the relationship is simple dominance?

2. The *no fangs* allele is **dominant** to the *fangs* allele (rather than the *no fangs* allele being recessive or incompletely dominant to the *fangs* allele).

   What is it about the **offspring data** that indicates that the *no fangs* allele is dominant to the *fangs* allele?

3. The gene for Fangs is **autosomal** (rather than X-linked).

   What is it about the **offspring data** that indicates that the Fangs gene is autosomal?

BEST COPY AVAILABLE

(Answer Key)

From Offspring to Mode of Inheritance: Worksheet Key

We often don't know the genotypes of individuals or the genetics of the species for a particular characteristic. One way to figure out the genetics of a particular characteristic is to carefully study of the patterns of inheritance of phenotypes.

Fangs

Another inherited characteristic in dragons is Fangs. Both Sandy and Pat have no fangs. But when you look at 100 of their offspring, you find the following:

- 29 (13 males and 16 females) have fangs
- 71 (37 males and 34 females) have no fangs

Monohybrid Inheritance III: Phenotypes to Genotypes

Use the information about the offspring to explain the mode of inheritance. Remember that in dragons, males are XX and females are XY.

3. The Fangs gene has two alleles–*fangs* and *no fangs*. The relationship between the two alleles is **simple dominance** (rather than incomplete dominance).

   What is it about the **offspring phenotypes** that indicates that the relationship is simple dominance?

   The relationship between the *fangs* and the *no-fangs* alleles is simple dominance because there are only two phenotypes among the offspring (fangs and no fangs).

3. The *no fangs* allele is **dominant** to the *fangs* allele (rather than the *no fangs* allele being recessive or incompletely dominant to the *fangs* allele).

   What is it about the **offspring data** that indicates that the *no fangs* allele is dominant to the *fangs* allele?

   The *no-fangs* allele is dominant to the *fangs* allele because approximately 3/4 of the offspring have the no-fangs phenotype and approximately 1/4 of the offspring have the fangs phenotype. Thus there is a 3:1 ratio of no fangs:fangs among the offspring.

3. The gene for Fangs is **autosomal** (rather than X-linked).

   What is it about the **offspring data** that indicates that the Fangs gene is autosomal?

   The gene for Fangs is autosomal because each phenotype among the offspring (fangs and no fangs) has approximately equal numbers of males and females.

BEST COPY AVAILABLE

(Problem Solution Explanation)

From Offspring to Modes of Inheritance: Teacher Information

This activity also deals with **monohybrid inheritance** but instead of going from genotypes to phenotypes or vice versa knowing the mode of inheritance, you need to figure out the mode of inheritance from parent and offspring phenotypes. In the fangs example, each parent has the no-fangs phenotype but some of their offspring have the fangs phenotype. To answer the three questions about fangs, you need to think about how the offspring data would look if the relationship between the two alleles was simple vs. incomplete dominance and if the Fangs gene was autosomal vs. X-linked. Since (1) both parents have the same phenotype and (2) some offspring have the same phenotype as the parents while some have a different phenotype, either both parents are heterozygous (if autosomal) or one parent is heterozygous (if X-linked).

	$\frac{1}{2}$ G	$\frac{1}{2}$ g
$\frac{1}{2}$ G	$\frac{1}{4}$ GG	$\frac{1}{4}$ Gg
$\frac{1}{2}$ g	$\frac{1}{4}$ Gg	$\frac{1}{4}$ gg

if autosomal

	$\frac{1}{2}$ G	$\frac{1}{2}$ g
$\frac{1}{2}$ G	$\frac{1}{4}$ GG	$\frac{1}{4}$ Gg
$\frac{1}{2}$ –	$\frac{1}{4}$ G–	$\frac{1}{4}$ g–

if X-linked

Let's use the two Punnett squares above to help think about the possibilities.

For Question 1, if the relationship between the two fangs alleles was **simple dominance**, then you would expect to see two phenotypes among the offspring,

- one phenotype corresponding to genotypes **GG** and **Gg** and a different phenotype corresponding to **gg** if the fangs gene was autosomal, or
- one phenotype corresponding to genotypes **GG**, **Gg**, and **G-** and a different phenotype corresponding to **g-** if the fangs gene was X-linked.

If the relationship between the two fangs alleles was **incomplete dominance**, then you would expect to see three phenotypes among the offspring,

- one phenotype corresponding to **GG**, one phenotype corresponding to **Gg**, and one phenotype corresponding to **gg** if the fangs gene was autosomal, or
- one phenotype corresponding to **GG** and **G-**, one phenotype corresponding to **Gg**, and one phenotype corresponding to **g-** if the fangs gene was X-linked.

Since there are only two phenotypes among the offspring, the dominance relationship between the two alleles for fangs must be **simple**.

For Question 2, given simple dominance, one of the fangs alleles must be dominant to the other. Among the offspring, you see approximately 3/4 with no fangs (the same as the parents) and 1/4 with fangs (different from the parents). The 3/4 no-fangs phenotype would correspond to the **GG** and **Gg** genotypes if the no-fangs allele (**G**) was dominant and the fangs gene was autosomal or to the **GG**, **Gg**, and **G-** genotypes if the no-fangs allele (**G**) was dominant and the fangs gene was X-linked. Thus the no-fangs allele (**G**) must be **dominant** to the fangs allele (**g**) and the fangs allele (**g**) must be **recessive** to the no-fangs allele (**G**).

For Question 3, you can distinguish between autosomal and X-linked inheritance by looking at the distribution of males and females for each phenotype among the offspring. If the Fangs gene was autosomal, you would expect each phenotypic class among the offspring to have approximately 1/2 females and 1/2 males. If the Fangs gene was X-linked, (a) males would be either **GG** or **Gg** so all males would necessarily have the no-fangs phenotype and (b) females would be either **G-** (no fangs) or **g-** (fangs). Thus, if the Fangs gene was X-linked, the fangs phenotype would have no males and the no-fangs phenotype would consist of 2/3 males and 1/3 females. Since the fangs and no-fangs phenotypes among the offspring have approximately equal numbers of males and females, the gene for Fangs must be **autosomal**.

BEST COPY AVAILABLE